Procesamiento del Lenguaje Natural, Revista nº 63, marzo de 2019, pp. 51-58 recibido 01-04-2019 revisado 09-05-2019 aceptado 14-05-2019

# Evaluation of Distributional Semantic Models for the Extraction of Semantic Relations for Named Rivers from a Small Specialized Corpus

## *Evaluación de Modelos Semánticos Distribucionales para la Extracción de Relaciones Semánticas Activadas por Ríos con Nombre Propio de un Corpus Especializado de Pequeño Tamaño*

**Juan Rojas Garcia, Pamela Faber**
Universidad de Granada
{juanrojas, pfaber}@ugr.es

**Abstract:** EcoLexicon (http://ecolexicon.ugr.es) is a terminological knowledge base on environmental science, whose design permits the geographic contextualization of data. For the geographic contextualization of landform concepts such as named rivers (e.g., *Nile River*), distributional semantic models (DSMs) were applied to a small-sized, specialized corpus to extract the terms related to each named river mentioned in it and their semantic relations. Since the construction of DSMs is highly parameterized and their evaluation in small specialized corpora has received little attention, this paper identified parameter combinations in DSMs suitable for the extraction of the semantic relations *takes_place_in*, *affects*, and *located_at*, frequently held by named rivers in the corpus. The models were thus evaluated using three gold standard datasets. The results showed that, for a small-sized corpus, count-based models outperformed prediction-based ones with the log-likelihood association measure, and the detection of a specific relation depended largely on the context window size.
**Keywords:** Named river, terminology, knowledge representation, distributional semantic model, text mining

**Resumen:** EcoLexicon (http://ecolexicon.ugr.es) es una base de conocimiento terminológica sobre el medioambiente, cuyo diseño permite la contextualización geográfica de los ríos con nombre propio (RNP) (v.gr., *Río Nilo*). Se aplicaron modelos semánticos distribucionales (MSD) a un corpus especializado de pequeño tamaño para extraer los términos relacionados con los RNP y sus relaciones semánticas. Puesto que el funcionamiento de los MSD depende de la configuración de sus parámetros, y su evaluación en corpus especializados de pequeño tamaño ha sido menos explorada, en este artículo se identifica la combinación de parámetros adecuada para extraer las relaciones semánticas *tiene_lugar_en*, *afecta* y *localizado_en*, activadas frecuentemente por los RNP. Los MSD se evalúan con tres conjuntos de datos anotados manualmente. Los resultados indican que, para un corpus de pequeño tamaño: los modelos basados en recuentos con la medida de asociación *log-likelihood* superan a los modelos predictivos; y la representación de una relación específica depende del tamaño de la ventana contextual.
**Palabras clave:** Río con nombre propio, terminología, representación del conocimiento, modelo semántico distribucional, minería de textos.

## *1 Introduction*

EcoLexicon (http://ecolexicon.ugr.es) is a multilingual, terminological knowledge base on environmental science that is the practical application of Frame-based Terminology (Faber, 2012). The flexible design of EcoLexicon permits the contextualization of data so that they are more relevant to specific subdomains, communicative situations, and geographic areas. However, the representation of geographically contextualized LANDFORM concepts, such as named rivers (e.g., *Mississippi River*), depends on knowing which terms are semantically related to each named

ISSN 1135-5948. DOI 10.26342/2019-63-5 

landform, and how these terms are related to each other.

With the aim of representing in EcoLexicon the conceptual structures underlying the usage of named landforms mentioned in a small-sized, English specialized corpus on Costal Engineering (7 million tokens), the terms related to each named landform and their semantic relations were extracted with distributional semantic models (DSMs) and other statistical techniques (Rojas Garcia and Faber, 2019a and 2019b). In this paper, we focus on named rivers.

DSMs represent the meaning of a term as a vector by considering the statistics of its cooccurrence with other terms in the corpus. Although a DSM can help identify semantic relations between terms, the construction of a suitable DSM for a particular task is highly parameterised. Even though numerous studies have addressed the evaluation and optimization of DSMs in very large, general corpora (Baroni et al., 2014; Kiela and Clark, 2014), the ability of DSMs to capture different semantic relations in smaller specialized corpora has received little attention.

The objective of this paper was to identify parameter combinations in DSMs suitable for the extraction of three semantic relations, held by named rivers, in the small specialized corpus mentioned above. Hence, the models were evaluated using evaluation data that contained pairs of semantically related terms, manually extracted from the same corpus. One of the terms was always a named river, and the other one was an entity or process. The semantic relations that linked the terms were those frequently activated by named rivers in the corpus, namely: (1) *takes_place_in*; (2) *affects*; and (3) *located_at*. Three gold standard datasets were thus built.

The rest of this paper is organized as follows. Section 2 provides background on DSMs, as well as a literature review on their application and evaluation. Section 3 explains the materials, methods, and DSMs evaluation applied in this study, and the construction of the gold standard datasets. Section 4 shows the results obtained. Finally, Section 5 discusses the results, and presents the conclusions derived from this work as well as plans for future research.

## 2 Background and Literature Review

Distributional semantic models (DSMs) represent the meaning of a term as a vector, based on its statistical co-occurrence with other terms in the corpus. According to the distributional hypothesis, semantically similar terms tend to have similar contextual distributions (Miller. and Charles, 1991). The semantic relatedness of two terms is estimated by calculating a similarity measure of their vectors, such as Euclidean distance or cosine similarity, *inter alia*.

Depending on the language model (Baroni et al., 2014), DSMs are either count-based or prediction-based. Count-based DSMs calculate the frequency of terms within a term's context (i.e., a sentence, paragraph, document, or a sliding context window spanning a given number of terms on either side of the target term). Correlated Occurrence Analogue to Lexical Semantic (COALS) (Rohde et al., 2006) is an example of this type of model.

Prediction-based models exploit neural probabilistic language models, which represent terms by predicting the next term on the basis of previous terms. Examples of predictive models include continuous bag-of-words (CBOW) and skip-gram models (Mikolov et al., 2013).

Count-based DSMs have been extensively studied (Kiela and Clark, 2014; Lapesa et al., 2014; Sahlgren and Lenci, 2016). Research shows that parameters, such as the context window size, influence the semantic relations that are captured, either syntagmatic relations or paradigmatic relations (i.e., synonymy, antonymy, hyponymy and meronymy). The syntagmatic relations examined in much research are either phrasal associates (e.g., *help - wanted*) (Lapesa et al., 2014) or syntagmatic predicate preferences (Erk et al., 2010) in general language. In this study, we focused on three specific syntagmatic relations, namely, *takes_place_in*, *located_at*, and *affects*, which were activated by named rivers in the specialized language of Coastal Engineering. As far as we know, this framework has not been studied in the context of DSM evaluation, which constitutes an original aspect of this work.

The ability of count-based models and prediction-based models (CBOW and skip-gram) to detect the three syntagmatic relations is described in this paper. Both types of DSM have also been recently compared. Baroni et al. (2014) compared them on several datasets and found that the prediction-based models provided better results. In contrast, Ferret (2015) found that count-based models performed better. In another work that compared the ability of both DSMs to capture paradigmatic relations (synonymy, antonymy, and hyponymy) and syntactic derivatives, Bernier-Colborne and Drouin (2016) observed that not only the semantic relations detected by the DSMs

depended on the window size, but also the values of this parameter mostly coincided in both DSMs.

Levy et al. (2015) yielded valuable insights, showing the following: (1) when the parameters of the models were tuned correctly, count-based and prediction-based models obtained similar accuracy; and (2) the best model depended on the task to be carried out. Nevertheless, Asr et al. (2016), and Sahlgren and Lenci (2016) reported that count-based models outperformed prediction-based ones on small-sized corpora of under 10 million tokens.

Work in lexical semantics and DSMs includes, *inter alia*, the identification of semantic relations (Bertels and Speelman, 2014), classification of verbs into semantic groups (Gries and Stefanowitsch, 2010), and the use of word vectors as features for automatic recognition of named entities in text corpora (El Bazi and Laachfoubi, 2016).

## *3 Materials and Methods*

### 3.1 Materials

#### 3.1.1 Corpus Data

The named rivers and related terms were extracted from a subcorpus of English texts on Coastal Engineering, on which the DSMs were also built. This subcorpus, comprising roughly 7 million tokens, is composed of specialized and semi-specialized texts, and is an integral part of the EcoLexicon English Corpus (23.1 million tokens).

#### 3.1.2 GeoNames Geographic Database

The automatic detection of the named rivers in the corpus was performed with a GeoNames database dump. GeoNames (http://www.geonames.org) has over 10 million proper names for 645 different geographic entities, such as bays, beaches, rivers, mountains, etc. For each entity, information about their normalized designations, alternate designations, latitude, longitude, and location name is stored.

#### 3.1.3 Gold Standard Datasets

The DSMs, built on our corpus, were evaluated on gold standard data, manually extracted from the same corpus. The gold standard datasets contained pairs of semantically related terms, in which the semantic relations were those frequently activated by named rivers in the corpus, namely, *takes_place_in*, *affects*, and *located_at*. Three gold standard datasets[1] were thus built, one for each of the semantic relations. It is necessary to note that the designations and meaning of these relations are those used in EcoLexicon (Faber et al., 2009).

The three semantic relations always linked a named river to an entity or process expressed by a nominal term. Specifically, the *takes_place_in* relation holds between a process (e.g., *runoff*) and a named river where the process occurs (see Table 1). The *affects* relation relates a named river to an entity or process (e.g., *soft mud*, *supply of sediment*) on which the named river performs a causative action (see Table 2). Finally, the *located_at* relation indicates the location of an entity (e.g., *jetty*) in a named river (see Table 3).

The three datasets contain 100 instances for each semantic relation, which were all used for the evaluation.

| process | *takes_place_in* | named river |
|---|---|---|
| consolidation of the land | *takes_place_in* | Mississippi river mouth |
| runoff | *takes_place_in* | Mississippi river basin |
| sea level rise | *takes_place_in* | Mississippi river delta |

Example from the corpus:
(1) ***Consolidation of the land*** *is occurring, as noted before, at the* ***mouth of the Mississippi River****, where the...*

Table 1: Extract from the gold standard dataset for the *takes_place_in* relation

| named river | *affects* | entity / process |
|---|---|---|
| Mississippi river | *affects* | Saint Bernard river delta |
| Mississippi river | *affects* | soft mud |
| Seine River | *affects* | fresh water |

Example from the corpus:
(1) *The Chandeleurs Islands are remnants of the* ***Saint Bernard River delta****, formed by the* ***Mississippi River****.*

Table 2: Extract from the gold standard dataset for the *affects* relation

| entity | *located_at* | named river |
|---|---|---|
| jetty | *located_at* | Mississippi river mouth |
| soft mud | *located_at* | Mississippi river mouth |
| barrier island | *located_at* | Mississippi river mouth |

Example from the corpus:
(1) *Natural and anthropogenic effects combine to result in the maximum erosional stress on* ***barrier islands****, located near the* ***mouth of the Mississippi River****.*

Table 3: Extract from the gold standard dataset for the *located_at* relation

The annotation of the pair of terms extracted from the corpus was carried out by three terminologists from our research group. *Cohen's*

[1] The datasets will be available on the website of the LexiCon Research Group at the University of Granada (http://lexicon.ugr.es/).

*kappa* coefficient was used as the statistical measure of inter-annotation agreement, and the scores for all the annotator pairs were over 90%.

## 3.2 Methodology

### 3.2.1 Pre-processing

The corpus texts were tokenized, tagged with parts of speech, lemmatized, and lowercased with the Stanford *CoreNLP* package (Manning et al., 2014) for R programming language (R Core Team 2019). The multiword terms stored in EcoLexicon were then automatically matched in the lemmatized corpus and joined with underscores.

In the DSMs, only terms larger than two characters were considered. Numbers, symbols, and punctuation marks were removed. Since closed-class words are often considered too uninformative to be suitable context words, stopwords were not used (i.e., determiners, conjunctions, relative adverbs, and prepositions). Additionally, the minimal occurrence frequency was set to 5 so that the co-occurrences were statistically reliable (Evert S., 2008).

### 3.2.2 Named River Recognition

Both normalized and alternate names of the rivers in GeoNames were searched in the lemmatized corpus. The recognized designations were normalized and automatically joined with underscores. Most rivers cited in the corpus were in GeoNames (97%), while others were identified by manual inspection (3%). Anaphoric elements referring to a river were replaced by the corresponding named river in the lemmatized corpus. For this task, the automatic anaphora resolution function from *CoreNLP* package was used, and other cases were manually replaced. The 250 rivers with the highest number of mentions in the corpus are shown on the map in Figure 1.

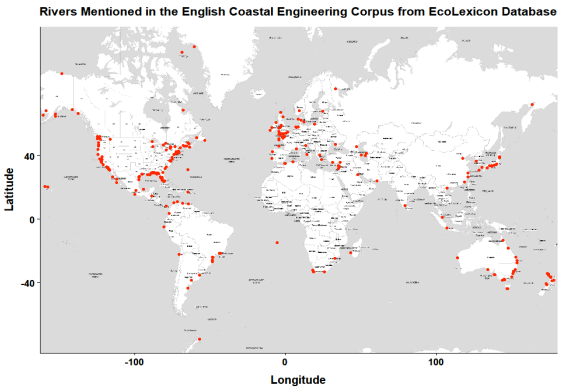


Figure 1: Location of the named rivers mentioned in the corpus

### 3.2.3 Construction of the DSMs

The experiment we carried out involved a comparative evaluation of two types of DSMs for a small-sized, specialized corpus, namely, count-based and prediction-based models. Both model types produced the vector representation of a term based on the contexts in which it appeared in our corpus. For this paper, the contexts of a target term (i.e., a named river) were the terms that co-occurred with it inside a sliding context window, which spanned a certain number of terms on either side of the target term.

These DSMs have various parameters that must be set to build the model. The parameters impinge on both the term representations produced, and the accuracy of the similarity scores between term vectors when the models are compared (Baroni et al., 2014).

Therefore, to assess the influence of the parameters of both DSMs on their ability to capture the three semantic relations targeted in this paper, various settings for each parameter were tried, and the combinations of these parameter settings were evaluated.

### 3.2.4 Parameter Setting of the Count-based Model

The first model type evaluated was a count-based model, also called bag-of-words (BOW) model. The BOW model was built with the R package *quanteda* (Benoit et al., 2018) for text mining.

To build a BOW model, a term-term matrix of co-occurrence frequencies was first computed, according to a specific size for the sliding context window. Then, the matrix was subjected to a specific weighting scheme, namely, an association measure that increases the importance of the context terms that are more indicative of the meaning of the target term. A dimensionality reduction technique could also transform this weighted matrix, but for this work, it was not applied. Therefore, the 3,000 most frequent words were used, which included all the named rivers and terms stored in the three evaluation datasets.

Regarding the context window, we tested size values ranging from 1 to 10 words on either side of the target term, and the context window was allowed to span sentence boundaries. The context window shape was always rectangular (i.e., the increment added to the co-occurrence frequency of a pair of terms was always 1, regardless of the distance between the two terms inside the context window). The frequencies observed on the left and right of a target term were added.

With respect to the weighting schemes, three association measures, defined in Evert S.'s (2008) work on collocation, were tested: (1) statistical log-likelihood; (2) positive pointwise mutual information (PPMI); (3) *t*-score. Log-likelihood and PPMI are widely used in computational linguistics, whereas *t*-score is popular in computational lexicography (Evert S. et al., 2017).

Research in computational linguistics reveals that log-likelihood is able to capture syntagmatic and paradigmatic relations (Lapesa et al., 2014), and to achieve better performance for medium-to-low-frequency data than other association measures (Alrabia et al., 2014). PPMI and *t*-score, on the other hand, have been found to work adequately for different applications in previous research when compared to other association measures (Baroni et al. 2014; Kiela and Clark, 2014).

Finally, following Lapesa et al. (2014), the association scores were transformed to reduce skewness in this way: log-likelihood and PPMI scores were both transformed by adding 1 and calculating then the natural logarithm (ln), whereas *t*-scores were transformed by calculating the square root (sqrt).

The settings tested for each of the two parameters were:

1. Size of the context window: 1-10 words.
2. Weighting scheme: ln(log-likelihood + 1), ln(PPMI + 1), sqrt(*t*-score).

#### 3.2.5 Parameter Setting of the Prediction-based Model

The second model type evaluated was a prediction-based model, specifically, the model known as *word2vec* (W2V) (Mikolov et al., 2013). These term vectors are learned by training a neural network on a corpus according to two different architectures. The architecture continuous bag-of-words (CBOW) aims to predict the target term based on its context terms, while the architecture skip-gram aims to predict the context terms of a target term. The W2V model was built with the R package *wordVectors* (Schmidt and Li, 2017).

For W2V, five hyperparameters were examined, the same as those tested by Bernier-Colborne G. and Drouin (2016) for paradigmatic relations and syntactic derivatives. The first one was the architecture used to learn the term vectors. The second one was the training algorithm, either using a hierarchical softmax function, or by sampling negative examples, in which case the number of negative samples must be selected. The third hyperparameter was the subsampling threshold for frequent terms, namely, some occurrences of those terms whose relative frequency in the corpus is greater than a threshold, are randomly deleted before the model is trained. Finally, the dimensionality of the term vectors, and the size of the context window are the other hyperparameters.

The settings tested for each of the five hyperparameters were:

1. Architecture: CBOW or skip-gram.
2. Negative samples: 5, 10 or none (in this case, hierarchical softmax is used).
3. Subsampling threshold: low ($10^{-5}$), high ($10^{-3}$) or none.
4. Dimensionality of term embeddings: 100 or 300.
5. Size of context window: 1-10 words.

#### 3.2.6 Evaluation of the DSMs

The DSMs were compared and described. First, for each named river, a sorted list of neighbours was obtained by computing the cosine similarity between the named river's vector and the vectors of all other context terms, and by sorting then these context terms in descending order of magnitude.

Subsequently, the sorted list of neighbours was evaluated on the whole gold standard dataset for each of the three semantic relations. The measure used to evaluate the models was mean average precision (MAP) (Manning et al., 1998). This measure tells us how accurate the sorted list of neighbours we get for a named river is, based on the rank of its related terms according to the gold standard. The nearer the related terms are to the top of this list on average for each named river, the higher the MAP.

## *4 Results*

BOW and W2V models were compared by observing the MAP of each model on the three datasets. The maximum MAP achieved by each model is shown in Table 4.

The results indicated, on the one hand, that the BOW model reached a higher MAP than W2V on the three semantic relations when its parameters were correctly tuned. On the other hand, the *takes_place_in* relation was the most accurately captured by both models when they were tuned for this relation, followed by the *located_at* and *affects* relations.

| Dataset | BOW model | | |
|---|---|---|---|
| | Max. MAP | Weighting scheme | Window size |
| *takes_place_in* | **0.544** (0.347 ± 0.118) | LL | 4 |
| *located_at* | **0.418** (0.321 ± 0.056) | LL | 2 |
| *affects* | **0.351** (0.201 ± 0.053) | LL | 3 |
| **Dataset** | **W2V model** | | |
| | **Max. MAP** | | **Window size** |
| *takes_place_in* | 0.346 (0.298 ± 0.042) | | 4 |
| *located_at* | 0.221 (0.196 ± 0.013) | | 2 |
| *affects* | 0.182 (0.141 ± 0.021) | | 3 |

Table 4: Maximum MAP (with average and standard deviation in brackets) of BOW and W2W models on each dataset. LL stands for the log-likelihood weighting scheme

The higher accuracy of *takes_place_in* may be due to the large number of instances in specialized texts in Coastal Engineering which express the processes that occur in named rivers. As for the *located_at* relation, it is also frequent that the texts mention the entities in named rivers. However, it seems that the number of instances of this semantic relation in the whole corpus is not large enough for the DSMs to represent them as accurately as *takes_place_in* instances.

We hypothesize that the low accuracy of both models to capture the *affects* relation could be caused by the lack of semantic expressivity of this relation, and by its high combinatorial potential to relate named rivers to both processes and entities. For instance, the conceptual proposition MISSISSIPPI RIVER *affects* SAINT BERNARD RIVER DELTA would be more meaningful if the relation were *creates* instead of *affects* (see sentence (1) in Table 2). The expressiveness of the *affects* relation could be increased by splitting it up into a number of different relations, as proposed by Reimerink A. and León-Araúz, (2017). In doing so, possibly, each of those specific relations could be more accurately represented by the DSMs.

Refining the *affects* relation not only would enhance the performance of DSMs in our opinion, but also would greatly improve the knowledge representation of named rivers, in the same way that the expressiveness of meronymy has been increased by splitting it up into six different relations in EcoLexicon (Reimerink and León-Araúz P, 2017).

Interestingly, in each dataset, the maximum MAP of W2V model was reached when the window size was the same as that of the BOW model. For that reason, to assess the impact of the window size on the accuracy of both DSMs, the average MAP for each setting of this parameter (i.e., for each window size between 1 and 10 words) is shown in Figure 2. The average MAP was used, instead of the maximum, because it allowed us to determine which window size settings consistently produced satisfactory results, regardless of the settings used for the other parameters.

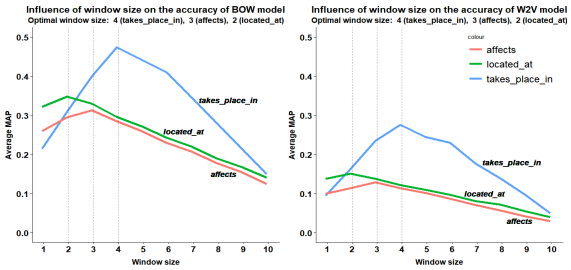


Figure 2: Average MAP of BOW model (left) and W2V model (right) with regard to window size

Figure 2 points out that, in both DSMs, the optimal size was 4 words for the *takes_place_in* relation, 3 words for *affects*, and 2 words for *located_at*.

Since the count-based model BOW notably outperformed the prediction-based W2V on the three datasets, for the sake of simplicity, the setting influence of the other four hyperparameters of W2V are succinctly reported because they did not lead to substantial accuracy improvements on either dataset. The neural network architecture skip-gram worked, on average, better than CBOW, and a negative sampling of 10 samples reached a larger MAP than the hierarchical softmax. The subsampling threshold was not conducive significant gains, and the optimal setting for the dimensionality of the term embeddings was 300 dimensions.

## 5 Conclusions

To extract knowledge for the representation in EcoLexicon of the conceptual structures (Faber P., 2012) that underlie the usage of named rivers in a small-sized, English Coastal Engineering corpus, count-based and prediction-based DSMs were applied to the corpus to extract the terms related to each named river. Since the construction of DSMs is highly parameterized, and their evaluation in small specialized corpora has received little attention, this paper identified parameter combinations in DSMs suitable for the extraction of the semantic relations *takes_place_in*, *affects*, and *located_at*, frequently held by named rivers in the corpus. The models were thus evaluated using three gold standard datasets.

Count-based models, with the log-likelihood association measure, showed the best performance for the three semantic relations. These results reinforce the findings of previous research which states, on the one hand, that count-based DSMs surpass prediction-based ones on small-sized corpora of under 10 million tokens (Asr et al., 2016; Sahlgren and Lenci, 2016), and on the other hand, that log-likelihood achieves greater accuracy for medium-to-low-frequency data than other association measures (Alrabia et al., 2014).

For both DSMs, the optimal window size depended on the semantic relation that was to be captured, and the specific values coincided in both DSMs, namely, a window size of 4 words for the *takes_place_in* relation, 3 words for *affects*, and 2 words for *located_at*. The dependence of the window size on the specific semantic relation, and the coincidence of the values in both DSMs are findings in line with those reported by Bernier-Colborne and Drouin (2016), upon comparing the ability of both DSMs to capture paradigmatic relations (synonymy, antonymy, and hyponymy), and syntactic derivatives.

It was also found that the *takes_place_in* relation was the most accurately represented by both DSMs, followed by *located_at*. The *affects* relation showed the lowest accuracy values, maybe due to the lack of semantic expressivity of this relation, and its high combinatorial potential to relate named rivers to both processes and entities (Faber et al., 2009). In future research, the *affects* relation will be split up into different relations, specific for named rivers, by following the methodology proposed by Reimerink A. and Léon-Araúz (2017), and by classifying the verbs that collocate with named rivers into the lexical domains and subdomains developed by Faber and Mairal (2009). In doing so, it will be analysed whether the DSMs reach higher accuracy values for each of the specific relations than for the *affects* relation.

Extensions of this work will include testing: other DSMs, such as GloVe; other parameters, such as the shape of the context window; and the application of the dimensionality reduction technique for texts Topic Modeling (Blei et al., 2003). Furthermore, the DSMs will be evaluated on gold standard datasets for named bays and beaches, and the three datasets for named rivers will be increased with more annotated data.

## Acknowledgements

This research was carried out as part of project FFI2017-89127-P, Translation-Oriented Terminology Tools for Environmental Texts (TOTEM), funded by the Spanish Ministry of Economy and Competitiveness. Funding was also provided by an FPU grant given by the Spanish Ministry of Education to the first author.

## References

Alrabia, M., N. Alhelewh, A. Al-Salman, and E. Atwell. 2014. An empirical study on the Holy Quran based on a large classical Arabic corpus. *International Journal of Computational Linguistics*, 5(1): 1-13.

Asr, F., J. Willits, and M. Jones. 2016. Comparing predictive and co-occurrence based models of lexical semantics trained on child-directed speech. In A. Papafragou, D. Grodner, D. Mirman and J. Trueswell (eds.) *Proceedings of the 38th Annual Conference of the Cognitive Science Society (CogSci)*, Philadelphia, 1092-1097.

Baroni, M., G. Dinu, and G. Kruszewski. 2014. Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the ACL*, vol. 1, 238-247.

Benoit K, K. Watanabe, H. Wang, P. Nulty, A. Obeng, S. Müller, & A. Matsuo. 2018. quanteda: An R package for the quantitative analysis of textual data. *Journal of Open Source Software* 3(30), 774.

Bernier-Colborne, G., and P. Drouin. 2016. Evaluation of distributional semantic models: a holistic approach. In *Proceedings of the 5th International Workshop on Computational Terminology (Computerm)*, Osaka (Japan), 52-61.

Bertels, A., and D. Speelman. 2014. Clustering for semantic purposes: Exploration of semantic similarity in a technical corpus. *Terminology*, 20(2): 279-303.

Blei, D.M., A.Y. Ng, and M.I. Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3: 993-1022.

El Bazi, I., and N. Laachfoubi. 2016. Arabic named entity recognition using word representations. *International Journal of Computer Science and Information Security*, 14(8): 956-965.

Erk, K., S. Padó, and U. Padó. 2010. A flexible, corpus-driven model of regular and inverse

selectional preferences. *Computational Linguistics*, 36(4): 723-763.

Evert, S. 2008. Corpora and collocations. In A. Lüdeling and M. Kytö (eds.), *Corpus Linguistics. An International Handbook*. Berlin: Walter de Gruyter, 1212-1248.

Evert, S., P. Uhrig, S. Bartsch, and T. Proisl. 2017. E-VIEW-alation – A large-scale evaluation study of association measures for collocation identification. In *Proceedings of the eLex 2017 Conference*, Leiden, 531-549.

Faber, P. (ed.). 2012. *A Cognitive Linguistics View of Terminology and Specialized Language*. Berlin/Boston: De Gruyter Mouton.

Faber, P., and R. Mairal. 2009. *Constructing a Lexicon of English Verbs*. Berlin/New York: Mouton de Gruyter.

Faber, P., P. León-Araúz, and J.A. Prieto. 2009. Semantic relations, dynamicity, and terminological knowledge bases. *Current Issues in Language Studies* 1: 1-23.

Ferret, O. 2015. Réordonnancer des thésaurus distributionnels en combinant différents critères. *TAL*, 56(2): 21-49.

Gries, S., and A. Stefanowitsch. 2010. Cluster analysis and the identification of collexeme classes. In S. Rice, and J. Newman (eds.), *Empirical and Experimental Methods in Cognitive/Functional Research*. Stanford (California): CSLI, 73-90.

Kiela, D., and S. Clark. 2014. A systematic study of semantic vector space model parameters. In *Proceedings of the 2nd Workshop on Continuous Vector Space Models and their Compositionality (CVSC)*, Gothenburg, 21-30.

Lapesa, G., S. Evert, and S. Schulte im Walde. 2014. Contrasting syntagmatic and paradigmatic relations: Insights from distributional semantic models. In *Proceedings of the 3rd Joint Conference on Lexical and Computational Semantics*, Dublin, 160-170.

Levy, O., Y. Goldberg, and I. Dagan. 2015. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, (3): 211-225.

Manning, C.D., P. Raghavan, and H. Schütze. 1998. *Introduction to Information Retrieval*. Cambridge (England): Cambridge University Press.

Manning, C.D., M. Surdeanu, J. Bauer, J. Finkel, S. Bethard, and D. McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, Baltimore, 55-60.

Mikolov, T., K. Chen, G. Corrado, and J. Dean. 2013. Efficient estimation of word representations in vector space. In *Workshop Proceedings of International Conference on Learning Representations*. Scottsdale.

Miller, G.A., and W.G. Charles. 1991. Contextual correlates of semantic similarity. *Language and Cognitive Processes* 6(1): 1-28.

R Core Team. 2019. *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.

Reimerink, A., P. and León-Araúz. 2017. Predicate-argument analysis to build a phraseology module and to increase conceptual relation expressiveness. In R. Mitkov (ed.), *Computational and Corpus-Based Phraseology*. Cham (Switzerland): Springer, 176-190.

Rohde, D., L. Gonnerman, and D. Plaut. 2006. An improved model of semantic similarity based on lexical co-occurrence. *Communications of the ACM*, 8: 627-633.

Rojas-Garcia J., and P. Faber. 2019a. Extraction of terms for the construction of semantic frames for named bays. *Argentinian Journal of Applied Linguistics* 7(1): 27-57.

Rojas-Garcia J., and P. Faber. 2019b. Extraction of Terms Related to Named Rivers. *Languages* 4(3): 46.

Sahlgren, M., and A. Lenci. 2016. The effects of data size and frequency range on distributional semantic models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, Austin (Texas), 975-980.

Schmidt, B., and J. Li. 2017. *wordVectors. Tools for creating and analyzing vector-space models of models of texts.* R package version 2.0.